Imputación de datos con operador OWA de la mayoría

Oscar A. Vallejos
Mgter. en Informática y Computación
Profesor Adjunto
ovallejos@exa.unne.edu.ar

María E. Valesani
Mgter. en Informática y Computación
Profesor Adjunto
evalesani@exa.unne.edu.ar

Osvaldo P. Quintana
Experto en Estadistica y Computación
Docente
oquin@mecon.indec.gov.ar

Universidad Nacional del Nordeste
Facultad de Ciencias Exactas y Naturales y Agrimensura
Dpto. de Informática
9 de Julio 1449
(3400) Corrientes, Argentina
(03783)-15-679884; (03783)-15-405117

Resumen

El presente trabajo tiene por objeto profundizar en la temática de la imputación de datos faltantes en base de datos con operadores de la mayoría MA-OWA, como técnica de imputación.
Se parte de una base de datos completa simulando pérdida de datos en distintos porcentuales. Se aplicó el operador OWA de la mayoría para su evaluación.
El trabajo se estructura de la siguiente manera: 1) introducción a la teoría de imputación de datos, 2) diseño del método MA OWA; 3) descripción del experimento y análisis de los resultados; 4) conclusión; 5) por ultimo se señala la referencia bibliografía que a permitido el conocimiento del arte al respecto.
Palabras claves: ma owa, imputación, operador de mayoría

Abstract

The present work takes as an object to study in depth the subject-matter of the imputation of data missing database with operators of the majority MA-OWA, as skill of imputation.
It breaks of a finished database simulating loss of information in different percentage. There applied the operator OWA of the majority for his evaluation.
The work is structured of the following way: 1) introduction to the theory of imputation of information; 2) design of the method MA OWA; 3) description of the experiment and analysis of the results; 4) conclusions; 5) finally it indicates to itself the reference bibliography that to allowed the knowledge of the art.

Word key: ma owa, imputation, operators of majority

1. Introducción a la imputación de datos

Es muy frecuente encontrar por diferente motivos, datos ausentes o *data missing* en bases de datos [3] [5].

La imputación de datos puede ser considera como la etapa final de un proceso de depuración de datos, ya sea por datos faltantes o valores cuyas reglas de edición han sido fallidas y serán reemplazados por valores aceptables conocidos.

La razón principal para realizar imputaciones es obtener un conjunto de datos completos y consistente al cual se le pueda aplicar las técnicas de estadística clásicas o inclusive se puede aplicar algún tipo de minería de datos.

Para la implementación de imputación de datos se recibe de la etapa anterior un fichero con ciertos campos marcados como "faltantes" ó "borrados" o *"fuzzy"* en la fase de edición por no cumplir algunas reglas propuestas. La imputación nos permitirá obtener un fichero completo. [5]

Encontrar el mejor método de imputación, o el más eficiente, es una tarea importante ya que se puede cometer errores en las imputaciones de datos individuales, e inclusive, pueden aparecer aumentados al realizar estadísticas agregadas. Por lo tanto se puede entender que es razonable estudiar métodos de imputación que conserven características de la variable como pueden ser: preservación de la distribución real del contenido de la variable, su relación con el resto de variables en estudio, etc.

A medida que crece el campo de la investigación sobre la temática de imputación es que se hace muy necesaria una comparación entre los distintos métodos vigentes en la actualidad. Este trabajo incorpora la técnica de operadores de agregación de la mayoría [4] como método de imputación.

El presente trabajo quiere tratar de acercar una medida de similitud entre la tabla original y la tabla imputada de forma tal de determinar la eficiencia de los operadores MA-OWA [2] [4].

2. Operadores MA-OWA

Una de esas variantes son los operadores OWA de la mayoría. Que evidentemente entendemos que puede ser un buen método para la imputación de datos faltantes y como mencionábamos en el resumen del presente trabajo, trataremos de demostrar, esta afirmación.

Como se muestra, el operador *MA-OWA* se define como:

$$F_{MA}(a_1.a_2.K.a_n) - \sum_{i=1}^{n} w_i.b_i - \sum_{i=1}^{n} f_1.(b_1.b_2.K.b_m).b_1$$

donde $w_i \in [0,1]$ con $\sum_{i=1}^{n} w_i = 1$ y b_i es el i-ésimo elemento de $a_1,\ldots, a_n$ ordenado en orden ascendente según las cardinalidades.

Los pesos del operador *MA-OWA* se calculan como sigue.

Sea δ_i la cardinalidad del elemento *i* con $\delta_i > 0$, entonces: $w_i = f_i(b_1.K.b_n)$ [2]

Los operadores de mayoría realizan la agregación en función del elemento δi que mide, de forma general, la importancia del elemento *i* a través de su cardinalidad. En los

procesos de mayoría se considera la formación de grupos de discusión o de mayoría en función de sus similitudes o distancias entre las opiniones de los expertos, considerándose dentro del mismo grupo todos aquellos valores con un mínimo de separación. Como se observa, la forma de cálculo del valor δ*i* es independiente de la definición de los operadores de mayoría. En este trabajo, la cardinalidad se calcula usando la siguiente función de distancia:

$$dist(a_i, a_j) = \begin{cases} 0 & \text{otro caso} \\ 1 & \text{if } \left|a_i - a_j\right| \le x \end{cases}$$

Por lo que la cardinalidad del elemento a_i es

$$\delta_i = \sum_{j=1}^{n} dist(a_i, a_j)$$

Donde el valor x modeliza el tamaño de cada grupo. Socialmente este grado se corresponde con la flexibilidad de los decisores para agruparse y reforzar sus opiniones.

Normalización de mayoría

En los operadores de mayoría cuantificada la cardinalidad total de los elementos y de cada uno de los grupos de mayorías se ve modificada por el cuantificador, por ello, los nuevos pesos cuantificados deberán modificarse para ser normalizados ya que se debe seguir cumpliendo la propiedad

$$\sum_{i=1}^{n} w_i = 1$$

Esta modificación se denomina normalización de mayoría y usa la suma de los valores del cuantificador como nueva cardinalidad, incrementando los pesos en la proporción expresada por dichos valores, de esta forma se obtiene una normalización cuantificada. La definición formal de la normalización de mayoría es la siguiente:

Sea $(w_1, w_2, ..., w_n)$ y $(q_1, q_2, ..., q_n)$ donde $0 \le q_i \le 1$ y, $\sum_{i=1}^{n} w_i = 1$ además se establece,

$$\rho = \sum_{i=1}^{n} w_i . q_i \le 1$$

Si $\varepsilon = 1 - \rho$, entonces $\varepsilon \ / \ \sum_{i=1}^{n} q_i$ es la cantidad a ser añadida a cada elemento en la cantidad expresada por q_i. Luego la normalización de mayoría es:

$$\sum_{i=1}^{n} (w_i . b_i + \frac{\varepsilon}{\sum_{i=1}^{n} q_i} . q_i) = 1$$

Como se observa en su definición, esta nueva normalización de mayoría modifica cada peso incrementándolo en función del cuantificador usado en cada caso con respecto a los demás elementos cuantificados, añadiendo, en función de este valor, a cada peso la

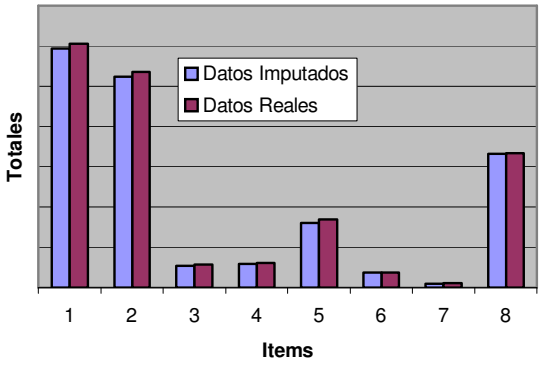
cantidad ε que corresponda. De esta forma se puede obtener una normalización que se integra en el proceso de agregación. La normalización clásica, por ejemplo, realiza el reajuste de los pesos considerando únicamente el tamaño del peso sin normalizar con respecto al tamaño total del conjunto que se desea normalizar, de esta forma se favorecen a aquellos elementos con mayor valor inicial en lugar de favorecer a aquellos que se han cuantificado más favorablemente (independientemente de su valor inicial).

3. Descripción del experimento

La base de datos tomada para el mencionado trabajo, tiene la característica general de poseer variables numéricas continuas y discretas. El tamaño de la muestra elegido para el experimento fue de 9 items o variables y 5.385 instancias. De los ítems, el 2, 3 y 9 corresponden a datos numéricos con decimales y son cifras mayores con respecto al resto. Los demás ítems corresponden a tipos de variables continuas.

Se practicó y registró la imputación a las bases de datos con distintos porcentajes de ausentismo. Se compararon y analizaron resultados, errores, distancias. [6]

Por la naturaleza de los datos, la imputación con este operador se comporto de forma mas que aceptable, tal como se muestra en la Figura 1, donde refleja el resultado de la imputación con el 5% de datos faltantes habiendo una distancia mínima entre los datos imputados y los datos reales.



Al conocer la filosofía del método (que tiene en cuenta la opinión de la mayoría y minoría), y dado que los datos tienen un rango de variación muy grande para algunos items, el MA OWA se comporto estable y con un error relativo aceptable.

Figura 1: Comparación de totales de datos reales e imputados por items para el 5% de datos faltantes

Se experimento la imputación con distintos porcentajes de datos faltantes del 5 al 30% [6]. Analizamos los resultados con el 30% y como refleja la Figura 2 el método se comportó con un error por debajo del 10% en el 40% de los item; y con un error aproximado del 20% en los ítems restantes. Se observó que el MA OWA para un porcentaje de datos faltantes elevado (30%), los errores son aceptables en relación al porcentaje de datos ausentes. El desvío estándar se mantuvo estable.

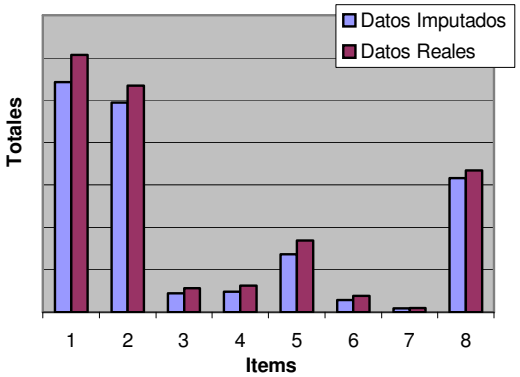


Figura 2: Comparación de totales de datos reales e imputados por items para el 30% de datos faltantes

En los casos de imputación numérica con decimales, el MA OWA alcanza su mejor performance.

5. Conclusiones

Como lo planteamos al principio, se hace difícil determinar cual es mejor método de imputación, o el más eficiente. Pero atendiendo a consideraciones que plantean diferentes autores sobre los métodos de imputación que conserven características de la variable como: preservación de la distribución real del contenido, su relación con el resto, tipo de variable que estemos tratando, comportamiento de los datos faltantes, etc., hemos concluido con un estudio exhaustivo sobre la imputación utilizando como método el operador MA-OWA.

Se ha incorporado, como un nuevo o posible método de imputación a este operador, dado que se ha comprobado que el método utilizado no ha generado errores adicionales, no se han alterados los números de atributos ni de instancias y la distribución de la frecuencia de las variables se mantuvo constante así como su desvió estándar.

El presente trabajo aporta un resultado exitoso de imputación con operador de la mayoría obteniendo un buen impacto en la calidad de los resultados. Habrá que continuar en esta línea de acción buscando la excelencia en métodos de imputación con otras técnicas como ser redes neuronales, algoritmos genéticos, u otras como es el estudio para ensayar la mejor representatividad de los operadores OWA y MA OWA, con la utilización de otros *ornes* para la obtención de los pesos.

5. Referencias

[1] J.I. Pelaez, J.M. Doña. Majority additive-ordered weighting averaging: a new neat ordered weighting averaging operators based on the majority process. *International Journal of Intelligent Systems*. 18(4):469- 481 (2003).

[2] J.I. Pelaez, J.M. Doña, D. La Red. Analysis of the majority process in group decision making process, in Proceedings 9th International Conference on Fuzzy Theory and Technology. North Carolina, USA. Pp.155- 159 (2003).

[3] Yang C. Yuan. *Multiple Imputation for Missing Data: Concepts and New Development*., SAS Institute Inc., Rockville, MD. P267-25

[4] Peláez J.I. Doña J.M. Mesas A. D. L. La Red *Opinión de mayoría en toma de decisión en grupo mediante el operador QMA-OWA*. Dpto. Lenguajes y Ciencias de la Computación E.T.S.I. Informática. Campus de Teatinos. Universidad de Málaga. 29071. Spain

[5] Little, R. & Rubin, D. B. (2002), Statistical Analysis with Missing Data, 2 edn, Jonh Wiley & Sons.

[6] Doña, J. M., Quintana, O. P., Valesani, M.E., Vallejos O.A. *Analysis of aggregation methods in incomplete database system* IPMU 2008